Capstone Project 2 - Netflix movies and TV shows clustering

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Individual Project.

Name : Shubham Chandrakar

Email-Id: chandrakar.shubham17@gmail.com

Contributions:

1) Data Wrangling
    a) Checking duplicate row
    b) Checking Null/Nan Values
    c) Feature engineering.
    d) Checking each individual feature.

2) EDA on Project
    a) Distribution of duration of content of TV series and Movies respectively.
    b) Top 5 directors on Netflix.
    c) Top 5 actors on Netflix.
    d) Top 10 Genres on Netflix.
    e) Distribution of ratings such as(PG-13, TV-MA etc.).
    f) Type of content released on Netflix after 2010.
    g) Sentiment analysis of contents in Netflix.
    h) Comparing count of movies and tv shows added since 2008
    i) The actor acted in the highest number of movies.

3) Data preparation for clustering
    a) Removing Punctuation
    b) Removing Stopwords
    c) Visualizing Word cloud of Description and genre
    d) Performing stemming

4) Topic Modeling and Clustering
    a) Preprocessing Data using Countvectorizer.
    b) Clustering using Latent Dirichlet Allocation (LDA).
    c) Clustering using K-means clustering algorithm.
    d) Clustering using Agglomerative clustering algorithm.

Please paste the GitHub Repo link.

Github Link :-
https://github.com/chandrakar-shubham/netflix-movies-and-tv-shows-clustering

Google drive link is posted below.

Google drive link :-
https://drive.google.com/drive/folders/148W57WlXMgyBj84pyGPZTAoUC5ucG9Jy?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The dataset provided to us by Flexible, a third party Netflix search engine. It consists of tv shows and movies available on netflix as of 2019. In 2019 Flexible released the report stating that the number of TV shows on Netflix has nearly tripled since 2010. While the number of movies on Netflix has decreased by 2,000 titles. So our aim is to perform exploratory data analysis (EDA), understand what type of content available in which country, to verify whether netflix is increasingly focusing on tv shows in spite of movie titles, and perform clustering of similar content by matching text based features.

As the first step, data wrangling is performed over raw data in the given database. After that we divided the project into various parts such as Exploratory data analysis (EDA) on content, Data preparation for clustering , Topic modeling and clustering.

In the Data wrangling step, individual analysis is performed over all features to analyze the distribution and trend of features. Null/Nan treatment is performed, Feature engineering is performed.

In the next step Exploratory Data Analysis (EDA) is performed. It is found that movie duration follows normal distribution and most of the TV Series had one season only, On average TV series have less than 5 seasons. Jan Sutler is top director and Anupam Kher is top actor in terms of number of involvement in various titles. Dramas, comedy, documentaries, action and adventure are the top four genres. Most of the titles have a TV-MA rating followed by a TV-14 rating. It is observed that after 2017 there is decline in addition of movie titles in comparison to TV series which are rising. In sentiment analysis of the description we can see that most of the content has positive sentiment. Various plots like box plot, line plot, scatter plot, count plot etc. is used to identify relationship between features

In the Data preparation step , We started with removing punctuations, followed by stopwords removal, then we performed stemming operations. Now data is ready for topic modeling.

In this step, we performed Topic modeling and Clustering, We first performed vectorisation using CountVectorizer(), then we used three clustering algorithms to find the optimum number of clusters. First we built the LDA model, these models after hyperparameter tuning give the result that the optimal number of clusters is two. Second model built was K-means, after evaluating the result of elbow method and silhouette score analysis it was found out that the optimum number of clusters is three. At last we built Hierarchical clustering model, after evaluating the dendrogram it was found out that the optimum number of clusters is 8.

Conclusion : It is found out that after 2017 there is decline in growth of new movie titles after 2017. Top 5 countries with the most movies on Netflix are the USA,India,UK, Canada and France . Top 5 countries with the most TV shows are the USA, UK, Japan, South Korea and Canada. As we performed topic modeling, LDA gives better results as data contains multiple topics. We can say that LDA with optimal number of clusters i.e. 2 is the best model.